图书情报工作
第 63 卷 第 13 期　2019 年 7 月

欧美科学数据开放存取出版平台服务调研及启示*

■ 秦顺　汪全莉　邢文明

湘潭大学公共管理学院　湘潭 411105

摘要：[**目的/意义**]数据的价值和科学数据开放存取出版的重要性已成为共识，欧美科学数据开放存取出版平台的建设经验具有借鉴意义。[**方法/过程**]选取欧美地区 14 个典型的科学数据开放存取出版平台为样本，根据科学数据出版“采集 – 分发 – 重用”的生命周期，从科学数据出版政策或愿景，科学数据整合、标识与交互，科学数据出版与分发，科学数据引用，数据生命周期管理与出版质量控制 5 个方面展开调研，归纳其服务建设特点与经验。[**结果/结论**]提炼得出对我国科学数据开放存取出版平台服务建设的有益启示：制定覆盖科学数据出版全生命周期的政策，重视科学数据出版服务建设的数据整合、数据标识、数据引用、数据评审等关键性问题。

关键词：科学数据　开放存取　数据出版　数据标识　数据引用

分类号：G237.9

DOI：10.13266/j.issn.0252 – 3116.2019.13.014

1　引言

科学数据是支撑科技创新、经济发展和国家安全的前提条件，具有较高的可重用价值，科学数据开放与出版受到各国际组织和国家的高度重视。欧盟地平线 2020 计划（Horizon 2020 programme）提出在欧洲的公共资助研究，要确保科学出版物的开放获取并且促进科学数据的开放获取，该计划资助的 FOSTER Plus 项目旨在促进地平线 2020 及往后开放科学的实际实施，指出“开放获取、开放数据”是开放科学运动的两支主要力量[1]。2004 年 OECD 发布的《开放获取公共资助研究数据的宣言》[2]及 2016 年欧盟研究与创新指导委员会出台的《OA 2020 计划行动纲要》[3]等皆将开放科学延伸至开放科学数据出版领域。我国政府也非常重视科学数据的共享与出版，2018 年 3 月国务院办公厅印发实施的《科学数据管理办法》第二十二条指出：“主管部门和法人单位应积极推动科学数据出版和传播工作，支持科研人员整理发表产权明晰、准确完整、共享价值高的科学数据。”[4]确立科学数据共享的公开出版模式，支持和推动科学数据惠及更广泛的科学研究领域是未来我国出版工作的重难点。

总体来说，前人对科学数据出版的研究主要围绕以下 4 个方面展开：①科学数据出版的动因。如 D. S. Sayogo 等[5]通过 DataONE 项目工作组进行的调查结果分析，指出数据共享与出版的动机主要包括数据管理技能和组织支持、通过法律政策等形成的对数据集创建者的认可机制两个关键因素。②科学数据出版模式。科学数据出版模式有两种、三种、四种和五种等不同划分方法，其中比较具有代表性的有：黄国彬[6]等归纳的科学数据集成出版、独立出版模式；涂志芳[7]总结的包括独立的数据出版、作为论文附件的数据出版和数据论文出版 3 种科学数据出版模式。③科学数据出版的关键性问题。学界比较重视从科学数据出版的单个环节或流程进行研究，亦有学者从数据生命周期理论考察分析，如涂志芳[7]认为数据标识、数据引用、数据评审为数据出版区别于一般数据共享的关键问题。④科学数据出版平台。其中大多数学者主要对科学数据出版平台的功能做出详细介绍，如 D. Roman 等[8]介绍了 DataGraft 提供的数据转换、发布和托管功能；王丹丹[9]结合对新加坡南洋理工大学科研人员的情景化访谈和 Dataverse 平台的使用测试分析，总结了科学数据出版平台的基本功能要求和用户体验要求；马建

* 本文系国家社会科学基金项目“信息生命周期视角下科研数据管理与共享的政策保障研究”（项目编号：15CTQ021）研究成果之一。

作者简介：秦顺（ORCID：0000 – 0003 – 1290 – 2266），硕士研究生，E-mail：1562131422@qq.com；汪全莉（ORCID：0000 – 0001 – 9007 – 8925），副教授，硕士生导师；邢文明（ORCID：0000 – 0001 – 8605 – 9107），副教授，硕士生导师。

收稿日期：2018 – 12 – 11　修回日期：2019 – 03 – 10　本文起止页码：129 – 136　本文责任编辑：徐健

玲[10]根据数据生命周期模型梳理了研究数据管理工具,其中涉及到了数据分发与出版类工具(平台);涂志芳等[11]亦梳理了部分科学数据开放存取出版平台及其功能。毋庸置疑,上述学者已从不同视角和维度对科学数据出版的认知性问题进行了阐述,可惜的是,鲜有针对科学数据开放存取出版平台服务建设的剖析和研究,也没有进行过整体性探讨。基于此,笔者通过对欧美典型的科学数据开放存取出版平台(以下简称"出版平台")服务调查和对比分析,试图明晰其整体服务现状和特点,探讨对于我国出版平台服务建设的启示。

2　调查对象和方法

2.1　调查对象

由于科学数据开放存取出版是较新的服务,国际上开展此服务的出版平台数量也较少。因此,为保证调查对象具有代表性,笔者设定了两个样本选取标准:①需满足科学数据开放存取(Open Access,OA)具备的开放性特征。出版平台应皆具有开源、开放的基本特点,具体表现为大多遵循开放存取的知识共享署名4.0国际许可协议(Creative Commons Attribution 4.0 International,CC BY 4.0)和 Github 开源支持,与本文调查指标高度相关。②组织者区域需具有代表性。欧美地区科学数据开放存取出版实践经验颇丰,长期处于领先地位,故选取组织者区域为欧美地区的出版平台。

综合以上标准,并通过参阅大量的科学数据出版主题文献和网络调研,依据平台的代表性、新颖性和资料的详尽程度,重点对比校正、选取了 14 个知名的出版平台作为调查对象,其组织者区域为欧美各占 7 个(见表 1),分别是:量化社会科学研究所(The Institute for Quantitative Social Science,IQSS)与哈佛大学图书馆、哈佛大学信息技术组织开发的 Dataverse,麻省理工学院图书馆和美国惠普公司实验室(Hewlett-Packard Labs)联合开发的 Dspace,美国国家科学基金会(National Science Foundation,NSF)资助 DataNet 计划研发的 DataONE,同是 NSF 资助的 DataNet 计划发起建立的 Data Conservancy,康奈尔大学 Albert R. Mann 图书馆运营的数据阶段型存储库 DataStaR,由 NSF 资助、vagrant-dryad 提供技术支持的 Dryad,斯坦福大学和 DuraSpace 合作开发的 Samvera(2017 年 5 月前被称为 Hydra 项目),非营利组织开放知识基金会(Open Knowledge Foundation,OKF)构建的 CKAN,CERN 数据中心提供技术支持的 CERN Open Data,EW-Shopp、proDataMarket 和 euBusinessGraph 项目运营的 DataGraft,由 Mark Hahnel 推出、Digital Science 支持的 Figshare,阿尔弗雷德韦格纳研究所、亥姆霍兹极地与海洋研究中心(The Alfred Wegener Institute,AWI)和不来梅大学海洋环境科学中心(Marine Umweltwissenschaften,MARUM)主办的 PANGAEA,Elsevier 开发的开放数据解决方案 Pure,由 CERN 数据中心和 OpenAIRE 提供支持的 Zenodo。

表 1　欧美科学数据开放存取出版平台基本情况

出版平台	区域	开发语言	开源支持	出版平台	区域	开发语言	开源支持
Dataverse	美国	HTML/Java	Github	CKAN	欧洲	HTML/Python/JavaScript	Github
Dspace	美国	HTML/Java	Github	CERN Open Data	欧洲	HTML/Python/JavaScript	Github
DataONE	美国	HTML/Java/Python	Github	DataGraft	欧洲	HTML/Java/Python	Github
Data Conservancy	美国	HTML/Java/Python	Github	Figshare	欧洲	HTML/PHP	Github
DataStaR	美国	HTML/Java	Github	PANGAEA	欧洲	HTML/R language	Github
Dryad	美国	HTML/Java/Python	Github	Pure	欧洲	HTML/Python	* *
Samvera	美国	HTML/Java	Github	Zenodo	欧洲	HTML/Python/JavaScript	GitHub

注:出版平台名称依组织者区域及英文字序排列,"* *"符号表示数据不详或无相关数据,下文相同

2.2　调查方法

根据科学数据出版"采集 - 分发 - 重用"的生命周期,拟定科学数据开放存取平台服务的调查指标共 5 个,分别是:①科学数据出版政策或愿景,解析出版平台科学数据开放存取相关政策;②科学数据整合、标识与交互,了解出版平台数据整合、标识与交互的工具与方法;③科学数据出版与分发,探索出版平台科学数据出版与分发路径、模式;④科学数据引用,梳理出版平台数据引用工具与方法;⑤数据生命周期管理与出版质量控制,探究出版平台中科学数据出版的生命周期完整性及出版质量控制。主要采用网络调查法,以深入使用平台的相关功能为前提,辅以文献调研逐一对各项指标内容进行归纳总结。本研究的调查统计时间为 2018 年 10 月 16 日至 2018 年 12 月 2 日。

3 调查结果分析

3.1 科学数据出版政策或愿景

科学数据的开放存取是大势所趋，英国工程和自然科学研究委员会(Engineering and Physical Sciences Research Council,EPSRC)指出:“不受限制地访问科学数据对于加速研究进展至关重要，科学和学术数据的数量每年呈指数级增长，但仍然缺乏利用这一重要资源的基础设施、政策与技术保障。”[12]建立科学数据开放存取出版平台是国内外推动数据开放的重要实践，而覆盖科学数据开放存取出版全生命周期的政策制定是其发展的先导。从调研情况来看(见表2)，所有的出版平台皆有推动科学数据共享出版和可重用的愿景，具有规范性政策的出版平台占比约为35.7%。Dataverse、Dryad、CERN Open Data、PANGAEA等平台制定了详尽的科学数据出版的相关规范、准则和条款，如Dryad推介了科学数据出版政策，包括Dryad数据出版内容标准、禁止发布数据的说明、撤回和删除数据的说明、提交者与使用者的权利和义务、隐私政策等[13]。

表2 开放存取出版重要政策或愿景调查

出版平台	类型	开放存取出版重要款目
Dataverse	政策	Dataverse社区规范、Harvard Dataverse一般使用条款、数据保存政策、数据隐私政策、Dataverse API使用条款、示例数据使用协议和复制数据集准则等[14]。
Dspace	愿景	提供使信息数据公开可用(可重用)和易于数据管理的手段[15]。
DataONE	愿景	开放、持久、稳健和安全地访问描述良好且易于发现的地球观测数据[16]。
Data Conservancy	愿景	数据保存、分享与发现:收集并处理研究数据，揭示数据在许多学科中的潜力，促进重用和进行新的数据组合[17]。
DataStaR	愿景	支持研究协作和数据共享，助力数据出版与高质量元数据存档[18]。
Dryad	政策	制定了科学数据出版政策，包括Dryad数据出版内容标准、隐私政策的等[13]。
Samvera	愿景	实现一体多用(One Body,Many Heads)及数据浏览查询、互操作，数据提交与重用[19]。
CKAN	愿景	实现数据访问，提供简化发布、共享、查找和使用数据的工具[20]。
CERN Open Data	政策	遵循CERN开放数据使用条款和隐私政策[21]。
DataGraft	愿景	用于数据转换、数据发布、数据托管与数据访问[22]。
Figshare	愿景	向世界开放科学数据，流程涉及数据的上传、管理、共享、发布。
PANGAEA	政策	遵循欧洲委员会《地平线2020计划科学出版物和研究数据开放获取指南》、DFG《关于保护良好科学实践的建议》、OECD《公共资金资助的研究数据获取原则与指南》、FAIR《科学数据管理指导性原则》等[23]。
Pure	愿景	实现验证和认证数据、捕获和重用数据，监控研究资助生命周期等。
Zenodo	政策	同CERN Open Data，遵循CERN开放数据使用条款和隐私政策[21]。

3.2 科学数据整合、标识与交互

3.2.1 科学数据整合　通过Semantic Web、Xml等进行数据封装整合、语义关联是目前应用较多的方式。据调查，所有平台皆使用Xml进行数据封装整合，比较具有个性化的有DataONE使用EML标准(Ecological Metadata Language,EML)编辑、整合元数据，DataStaR以Semantic Web进行科学数据语义关联，Data Conservancy以RMap(关联数据图)、GUI方式进行数据关联和数据集封装，DataGraft明确使用RDF资源描述框架进行数据描述，PANGAEA和Zenodo则以OAI-PMH接口实现数据收割采集与整合，拓宽科学数据共享与出版范围。可见，语义关联化是科学数据整合的一大趋势。见表3。

3.2.2 科学数据标识　数据标识是数据出版、分发和引用的前提，同时亦作为数据封装整合、数据交互的枢纽。海量数据使得科学数据的定位与标识难度加大，故需对科学数据进行整合与标识。数字对象标识符(Digital Object Identifier,DOI)具有唯一且永久标识、永久定位等特点，适合且利于数据开放存取出版，赋予科学数据DOI号将伴随数据的整个出版过程[24]。由表3可知，14个出版平台中有10个使用DataCite、EZID、CrossRef等工具进行DOI注册，12个平台与ORCID Inc.合作赋予数据作者唯一标识ID，即开放研究员和贡献者ID。DOI和ORCID利于解决数字版权管理和知识产权问题，减少了数据交互纠纷。

3.2.3 科学数据交互　从表3可以看出，有11个出版平台以API方式进行科学数据交互，其中通过独具REST风格(REpresentational State Transfer，表现层状态转移)的API进行数据描述、数据交互是主要做法，RESTful API具有统一接口URI，能够基于HTTP协议实现多种格式的数据调用，极大地扩展了科学数据开放存取出版的覆盖面。此外，DataStaR等出版平台基于自主研发框架，创新了数据交互方式。但是，通过API方式进行科学数据标识、关联与交互是一个统一规范的路径。

表 3 科学数据整合、标识与交互方式调查

出版平台	数据整合方式	作者标识工具	数据标识工具	数据交互方式
Dataverse	Xml	ORCID	DataCite/EZID	API
Dspace	Xml	ORCID	CNRI	RESTful API
DataONE	Xml/EML	ORCID	DataCite/Morpho	API
Data Conservancy	Xml/RMap/GUI	* *	* *	API-XI
DataStaR	Xml/Semantic Web	ORCID	* *	Linked Data
Dryad	Xml	ORCID	DataCite	RESTful API
Samvera	Xml	ORCID	* *	API
CKAN	Xml	ORCID	EZID	API
CERN Open Data	Xml	ORCID	DataCite	RESTful API
DataGraft	Xml/RDF	* *	* *	API
Figshare	Xml	ORCID	DataCite/EZID	API
PANGAEA	Xml/OAI-PMH	ORCID	PANGAEA DOI	URL
Pure	Xml	ORCID	CrossRef	* *
Zenodo	Xml/OAI-PMH	ORCID	DataCite	RESTful API

综上所述,科学数据整合、标识与交互是数据生命周期视角下科学数据出版相互联系较为紧密的环节,是实现科学数据有序出版的核心业务工作流程。由调研分析可以看出,欧美出版平台中科学数据整合、标识与交互技术工具的开发应用独具个性化特征,但从某种程度上而言亦缺乏标准化、规范化引导,系统异构加之辅助出版工具多元化,很大程度上加大了科学数据整合、标识与交互工作的难度。

3.3 科学数据出版与分发

欧美出版平台的科学数据出版与分发成绩斐然,如 Dataverse 共收录了 52 449 个数据集(Datasets)、产生 4 102 900 次下载(Downloads)[25],Dataverse、Dspace 等开源框架被我国的北京大学、武汉大学等引入或合作开发了科学数据管理与出版平台可为佐证。据调研(见表 4),大多数的出版平台皆支持 zip、xlsx、csv 等格式,支持的科学数据出版格式具有多样化特征。其中以 Dryad 做得最为细致,其规定了首选格式和格式支持级别,包括文本、图像、音频、视频、压缩文档等类型数据的首选格式,并对这些数据类型划分了格式支持级别,即全力支持、有限支持和原始比特流访问 3 个级别[13]。

表 4 科学数据出版格式与共享协议调查

出版平台	主要数据格式	共享协议	出版平台	主要数据格式	共享协议
Dataverse	xlsx/csv/tsv	CC BY 4.0/CC0	CKAN	csv/pdf/doc/xlsx	AGPL
Dspace	pdf/doc/jpeg/tiff	CC BY 4.0/BSD	CERN Open Data	zip/root/gz/pdf	CC0/XROOT
DataONE	EML v2.1.0/pdf	* *	DataGraft	csv/rdf/xml/turtle	* *
Data Conservancy	* *	* *	Figshare	text/tgz/data	CC BY 4.0
DataStaR	csv/rdf	CC BY 4.0	PANGAEA	zip/txt/tab/jpeg	CC BY 3.0
Dryad	xlsx/csv/txt/tex	CC0/CC BY 3.0	Pure	zip/txt/xlsx	GDPR
Samvera	* *	CC BY 4.0/Apache 2.0	Zenodo	zip/nex/xlsx/odg	CCO

知识共享与知识产权保护是科学数据出版与内容分发面临的主要矛盾,而建立具有约束力的共享协议是解决这一矛盾的主要途径。欧美出版平台的科学数据出版大多遵循约定共享协议分发。由表 4 可见,有 11 个出版平台明确遵循 CC BY 4.0、CCO 等共享协议,重视科学数据出版与分发过程中的知识产权保护。不同共享协议各有千秋,遵循创作许可(CC)等知识共享协议提供的豁免与规范,能够保护使用或重新分发数据作者作品的科研人员免受版权侵权的关注,同时保护数据拥有者的权利[26],权衡版权保护与共享利用,使权责明晰。

3.4 科学数据引用

数据引用是数据出版的关键环节,是保障数据作者与管理者数据权益的一种有效方式[7]。Dataverse 等 6 个出版平台认可数据引用原则联合声明(FORCE11)及"可发现(Findable)、可访问(Accessible)、可互操作(Interoperable)和可重用(Reusable)"的 FAIR 数据共享原则[27],大多数的出版平台数据引用则使用 DOI 和

秦顺，汪全莉，邢文明.欧美科学数据开放存取出版平台服务调研及启示[J].图书情报工作,2019,63(13):129－136.

数据指纹技术(Universal Numeric Fingerprints,UNF),并遵循国际科技数据委员会(CODATA)、Datacite 等的引用原则,通过 Datacite、Crossref 等赋予科学数据 DOI,生成引文格式化程序(DOI Citation Formatter),能够支持不同的引用语言和引用风格[28]。目前支持基于 DOI 的科学数据引用工具主要为 Datacite、Mendeley、EZID、Zotero 等,这些工具具有良好的支持性,譬如 Zotero 适用于不同的数据格式,其插件支持 Firefox、Chrome 和 Safari 等端口,同时适用于 Word、LibreOffice、BibTeX 和 LaTeX 等文字处理软件[29]。出版平台常用的科学数据引用工具如图 1 所示:

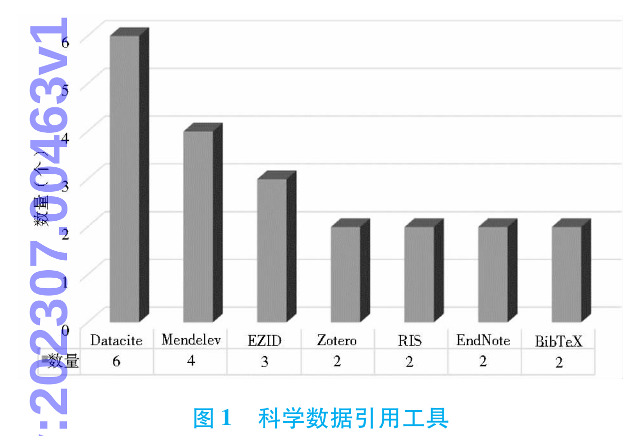


图 1　科学数据引用工具

3.5　数据生命周期管理与出版质量控制

数据生命周期管理是数据出版质量控制的必要条件,具备完整的数据出版生命周期管理的出版平台,其数据审查机制、数据质量控制亦相对完善。DataONE 将数据生命周期分为计划、收集、确保、描述、保留、发现、集成、分析 8 个组件[30],笔者在调研梳理 14 个出版平台的数据生命周期管理情况后,参照 DataONE 数据生命周期模型和 A. Sarretta 提出的《研究数据生命周期》(Research Data Life Cycle)[31],将科学数据出版的生命周期划分为采集(Collect)、标识(Identify)、出版(Publish)、分发(Distribute)、重用(Reuse)、评价(Evaluate)6 个阶段,建立的科学数据出版生命周期模型(Data Publishing Life Cycle,DPLC)见图 2。目前完整涉及科学数据出版生命周期的出版平台仍不多,尤其是在评价环节,出版质量控制方面仍亟待优化,评审与共享之间的平衡问题仍在进一步研究和实践之中。引入数据质量控制计划和同行评审制度将有利于数据出版质量控制,如 Dryad 承诺其工作人员与同行评议人员会在数据发布之前对数据的安全性、学术性、技术正确

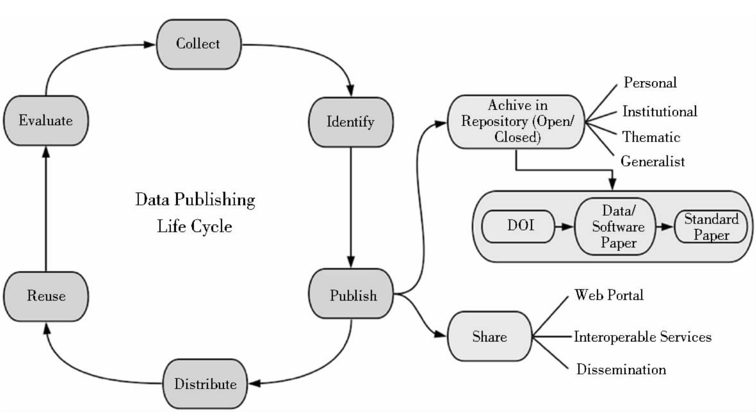


图 2　科学数据出版生命周期模型[31]

性等进行审查和管理,DataONE 亦从多个方面对数据出版进行质量审查,可资借鉴[32]。

4　对我国科学数据开放存取出版平台服务建设的启示

我国的科学数据开放存取出版平台建设起步较晚,相关服务亦相对落后,但已受到国家、科研教育机构等主体的高度重视,并在数据综合集成加工与发布、数据资源发现与检索、数据下载(重用)与共享等方面展开了有益探索,总体发展空间很大。截至目前,科技部、财政部先后在基础科学、农业、林业等 8 个领域建成了国家科技资源共享服务平台;中国科学院牵头建成国家基础科学数据共享服务平台,涵盖"一主一备＋12 分中心"分布式、可扩展存储系统;北京大学、武汉大学等则建设了高校系统的科学数据共享平台。但是,总体而言我国科学数据开放存取出版实践较为贫弱,表现为出版平台较少、相关工具匮乏以及标准规范、技术条件、人才队伍等不完善问题。

进一步改善我国出版平台的科学数据开放存取出版服务建设需要吸收欧美出版平台积累的经验:①欧美出版平台的优势。政策体系完善、服务意识较高、服务内容多样、服务共享程度高、标准体系健全、技术条件更为成熟等为可鉴之处。另值得指出的是,欧美出版平台尤为重视数据生命周期视角下的科学数据出版相关工具的开发和应用,较好地提升了科学数据开放存取出版质量和效率。②欧美出版平台的不足。存在个性化与标准化建设的偏离,个性多元的技术标准一定程度上限制了科学数据开放存取出版,数据整合、数据标识、数据引用和数据评审等关键问题的处理皆囿于各行其是的流程,较难实现异构整合。因此,数据出版标准、规范及其技术实现将成为未来研究与实践的

一大重点[7]。有鉴于此,我国的出版平台服务建设应采取积极的优化策略。

4.1 制定覆盖科学数据出版全生命周期的政策

4.1.1 标准化与规范化政策　实现开放获取、开放数据和开放科学,要进行科学数据开放共享各个环节的政策研究[33]。科学数据出版服务面临着复杂的知识产权问题,是一项繁杂的业务工作,为保证这项工作后顾无忧,制定和实施标准化与规范化政策显得尤为迫切。《科学数据管理办法》的印发施行,为科学数据共享与出版提供了指引,但其整体仍处于一个尚在探索的领域[34]。出版平台的服务建设要在探索科学数据集成出版、独立出版两种开放存取模式多样互补的同时,注重标准化与规范化建设。在科学数据出版环节,可于宏观层面加强对数据整合、数据出版服务、平台交流共享等内容的构建,制定涉及科学数据出版全生命周期的标准化与规范化政策,建立专门的实施小组或委员会,对出版平台数据出版工作进行指导与监管。相关出版平台亦需在以上基础上制定实施细则(如 Dryad 数据出版内容标准),实施细则应包含科学数据的采集、标识、出版、分发、重用、评价 6 个层面。此外,还需探索标准化与规范化的多方合作机制,解决出版平台在科学数据开放存取出版过程中的服务理念、技术方法等方面的对接冲突。

4.1.2 开源与自主研发驱动政策　细观欧美出版平台的建设,验证了开源与自主研发的强大优势,其中 13 个平台支持以 GitHub 托管和审查代码、管理项目和构建软件[35]。在开源环境下,开放存取政策驱动与标准化约束为解决我国科学数据出版基础贫弱之道。首先,制定开放存取政策,鼓励开源与自主研发应为重要着手点。科学数据的开放存取模式是:数据作者创作科学数据→存储于科学数据开放存取出版平台(数据知识库)→学者免费利用科学数据创造新的学术成果,该模式的运行需要政策、财政与技术保障,否则难以维持。其次,处理好个性化与标准化研发的关系。鼓励引进先进开源平台的同时,要注重标准化整合与自主研发能力的提升,如北京大学图书馆对结构化、半结构化和非结构化数据给予管理支持,采用 Dataverse 平台架构开发了学科开放数据导航[36],支持科学数据开放存取出版;中国科学院兰州文献情报中心自主研发的全球科研项目数据库(ProjectGate),提供数据提交、数据审核发布等服务[37]。再者,需注重对开源与自主研发驱动政策、出版平台使用方法等进行宣传,提升科学数据开放存取的实际效能。

4.2 重视科学数据出版服务建设的关键性问题

4.2.1 数据整合:促使数据序化组织　科学数据的序化整合是开放存取出版的前提条件。目前,数据整合面临的主要挑战有系统异构、科学数据描述语法不统一、科学数据元数据格式不统一、科学数据之间缺乏语义关联等[38]。可借鉴 Data Conservancy 构建的 4 个关键组件,分别是 DC 包装规范(DC Packaging Specification)、包摄取服务(Package Ingest Service)、关联数据图(RMap)和 API 扩展体系结构(APIX)[17],形成数据整合到数据出版的标准化、规范化流程。通过一站式数据整合,便于数据作者序化组织科学数据,促进数据重用与科学研究的推陈出新。

4.2.2 数据标识:赋予唯一永久标识　国内的科学数据出版主要为简易数据发布与共享,与具有"来源可靠、质量可信、公开发布、公共利用、唯一标识、知识产权清晰、可正式引用"[7]等特征的开放存取出版仍有一定的差距。就数据标识层面而言,需要进行赋予促进数据交互的唯一永久标识符,推动科学数据开放存取出版。具体需以数据标识与作者标识并重为策略,DOI 与 ORCID 皆具有唯一永久标识性,能确保科学数据与数据作者永久关联,是国外出版平台的实践热点,在我国科技论文标识领域亦有深层次应用。不同的是,对象标识符(Object Identifier,OID)为我国各科学系统规范采用的科学数据唯一永久标识符,由科学数据主管部门向国家 OID 注册中心申请获得[39]。因此,要深化 OID、DOI 与 ORCID 等标识工具的应用,为标准化科学数据引用做铺垫。

4.2.3 数据引用:保障知识产权清晰　14 个出版平台的开放存取服务实践表明,数据引用能够保障知识产权清晰,确保科学数据合理使用。因此,要加强科学数据引用标准的制定和科学数据引用工具的开发应用。2017 年 12 月,我国印发了《GB/T 35294 –2017 信息技术 科学数据引用》国家标准,通过"通用科学数据引用格式"和"基于 OID 的科学数据引用方式"规范科学数据引用,引用格式如下[39]:①通用科学数据引用格式:作者. 名称(版本). 创建机构[创建机构],创建时间. 传播机构[传播机构],传播时间. 唯一标识符;解析地址。②基于 OID 的科学数据引用方式:科学数据 OID 标识前缀. 出版厂商代码. 科学数据唯一代码。两种引用标准皆可厘清数据来源,确保科学数据的唯一性,便于声明科学数据传播的路径。我国的出版平台应在上述标准引导下,充分考虑科研人员的数据需求,开发适用多元数据格式和复杂端口的数据标识、数

据引用自动化生成工具,保障数据作者的知识产权,同时提升科学数据引用、传播的质量和效率。

4.2.4 数据评审:确保数据质量可靠 出版平台进行数据评审主要有出版平台的工作人员和同行评议人员两个主体,其中工作人员侧重于数据的技术质量审核,同行评议人员则偏向于关注数据的科学质量。技术质量审核的内容为数据完整性和描述的充分性,科学质量审核的主要内容包括数据完整性、描述的详细程度、数据有用性等[7]。科学数据同行评议是保证数据质量的重要手段之一,对于产生正确的科学结果有重要意义,一些工具和过程可能有助于快速、便捷地开展数据同行评议[40]。同行评议在出版平台数据评审中应用较为贫弱,可借鉴 Dataverse 等的实践,引入数据质量控制计划,如采用数据管理计划(Data Management Plan,DMP)与 DMPTool 来评估测试数据质量,为简化同行评议做准备[41]。此外,在出版平台数据评审过程中需要考虑智能化审查、工作人员质量控制和同行评议并行,监管与服务并重,有利于解决数据在语义、语用层面的评估,实现多学科、多领域的数据质量控制,这些问题还有待研究和实践。

5 结语

本研究基于科学数据出版“采集－分发－重用”的生命周期,综合运用文献研究、网络调研、对比分析等研究方法对 14 个科学数据开放存取出版平台的科学数据出版政策或愿景,科学数据整合、标识与交互,科学数据出版与分发,科学数据引用,数据生命周期管理与出版质量控制等服务内容进行梳理,认为其积累的经验可为我国出版平台服务建设在科学数据出版政策制定、关键业务工作发展过程中提供参考依据,希望能助推我国科学数据开放存取出版工作的持续健康发展。本文亦有不足之处,笔者仅选取欧美地区 14 个出版平台进行分析,样本数量略微偏少,在后续的研究中,有必要进一步扩大调查对象和调查范围,细化调查指标,提高研究结果的准确性和应用价值。

参考文献:

[1] FOSTER Plus. Open science[EB/OL]. [2018－11－02]. https://www.fosteropenscience.eu/foster#taxonomy.

[2] OECD. Declaration on access to research data from public fundinge[EB/OL]. [2018－11－02]. https://legalinstruments.oecd.org/en/instruments/157.

[3] EUROPEAN COMMISSION Directorate-General for Research/Innovation. Guidelines to the rules on open access to scientific publications and open access to research data in horizon 2020[EB/OL]. [2018－11－02]. https://legalinstruments.oecd.org/en/instruments/157.

[4] 国务院办公厅. 关于印发科学数据管理办法的通知[EB/OL]. [2018－11－03]. http://www.gov.cn/zhengce/content/2018-04/02/content_5279272.htm.

[5] SAYOGO D S, PARDO T A. Exploring the determinants of scientific data sharing: understanding the motivation to publish research data[J]. Government information quarterly, 2013, 30(S1): 19－31.

[6] 黄国彬,王舒,屈亚杰. 科学数据出版模式比较研究[J]. 大学图书馆学报,2018,36(1):34－40,33.

[7] 涂志芳. 科学数据出版的基础问题综述与关键问题识别[J]. 图书馆,2018(6):86－92,100.

[8] ROMAN D, DIMITROV M, NIKOLOV N, et al. Datagraft: simplifying open data publishing[C]//European Semantic Web Conference: The Semantic Web. Berlin: Springer, 2016: 101－106.

[9] 王丹丹. 科学数据出版平台的用户测试研究[J]. 情报资料工作,2017(6):56－61.

[10] 马建玲,曹月珍. 研究数据管理工具发展研究[J]. 图书馆学研究,2014(15):40－47.

[11] 涂志芳,刘兹恒. 国内外学术图书馆参与开放存取出版的实践进展述略——从出版途径的视角[J]. 图书与情报,2017(3):61－71.

[12] SPARC. Open data[EB/OL]. [2018－11－09]. https://sparcopen.org/open-data/.

[13] DATADRYAD. Terms of service[EB/OL]. [2018－11－04]. https://datadryad.org/pages/policies#publication.

[14] DATAVERSE. Dataverse community norms[EB/OL]. [2018－11－04]. https://dataverse.org/best-practices/dataverse-community-norms.

[15] DSpace. About DSpace[EB/OL]. [2018－11－04]. https://duraspace.org/dspace/about/.

[16] DataONE. What is DataONE? [EB/OL]. [2018－11－04]. https://www.dataone.org/what-dataone.

[17] Data conservancy. Building infrastructure for data curation[EB/OL]. [2018－11－04]. https://dataconservancy.org/.

[18] Researchgate. DataStaR: a data sharing and publication infrastructure to support research[EB/OL]. [2018－11－04]. https://www.researchgate.net/publication/267776401_DataStaR_A_Data_Sharing_and_Publication_Infrastructure_to_Support_Research.

[19] SAMVERA. How it works[EB/OL]. [2018－11－04]. https://samvera.org/.

[20] CKAN. CKAN, the world's leading open source data portal platform[EB/OL]. [2018－11－04]. https://ckan.org/.

[21] CERN. CERN open data terms of use[EB/OL]. [2018－11－05]. http://opendata.cern.ch/docs/terms-of-use.

[22] DATAGRAFT. About us, DataGraft development and funding[EB/OL]. [2018－11－04]. https://datagraft.io/about-us.

[23] PANGAEA. About[EB/OL]. [2018－11－05]. https://www.

pangaea. de/about/.
[24] 涂勇,彭洁. 数字对象唯一标识在中国科学数据领域中的应用研究[J]. 数字图书馆论坛,2013(8):31 - 36.
[25] DATAVERSE. Open source research data repository software[EB/OL]. [2018 - 11 - 10]. https://dataverse. org/.
[26] WIKIPEDIA. Creative Commons license[EB/OL]. [2018 - 11 - 10]. https://en. wikipedia. org/wiki/Creative_Commons_license.
[27] WILKINSON M D, DUMONTIER M, AALBERSBERG I J J, et al. The FAIR Guiding Principles for scientific data management and stewardship[EB/OL]. [2018 - 11 - 10]. https://www. nature. com/articles/sdata201618. pdf.
[28] DATACITE. Citation formatter[EB/OL]. [2018 - 11 - 11]. https://www. datacite. org/citation. html.
[29] MIT Libraries. Citation management and writing tools: citation management tools[EB/OL]. [2018 - 11 - 11]. https://libguides. mit. edu/cite-write/citetools.
[30] DataONE. Data life cycle[EB/OL]. [2018 - 11 - 11]. https://www. dataone. org/data-life-cycle.
[31] SARRETTA A. Research data life cycle[EB/OL]. [2018 - 11 - 11]. http://doi. org/10. 5281/zenodo. 1149049.
[32] 涂志芳,刘兹恒. 国外数据知识库模式的数据出版质量控制实践研究[J]. 图书馆建设,2018(3):5 - 13.
[33] 顾立平. 科研模式变革中的数据管理服务:实现开放获取、开放数据、开放科学的途径[J/OL]. 中国图书馆学报,2018(6):1 - 15. https://doi. org/10. 13530/j. cnki. jlis. 180018.
[34] 邢文明,洪程. 开放为常态,不开放为例外——解读《科学数据管理办法》中的科学数据共享与利用[J/OL]. 图书馆论坛,2019(1):1 - 8. http://kns. cnki. net/kcms/detail/44. 1306. G2. 20180818. 1836. 002. html.
[35] GitHub, Inc. Built for developers[EB/OL]. [2018 - 11 - 13]. https://github. com/.
[36] 北京大学图书馆. 学科开放数据导航[EB/OL]. [2018 - 11 - 14]. http://www. lib. pku. edu. cn/portal/cn/fw/sjfw/xuekeshuju.
[37] 中国科学院兰州文献情报中心. 全球科研项目数据库 Know what you want,offer what you need[EB/OL]. [2018 - 11 - 17]. http://project. llas. ac. cn/index. jsp.
[38] 白如江,冷伏海. "大数据"时代科学数据整合研究[J]. 情报理论与实践,2014,37(1):94 - 99.
[39] 国家质量监督检验检疫总局,国家标准化管理委员会. 信息技术 科学数据引用[EB/OL]. [2018 - 11 - 14]. http://c. gb688. cn/bzgk/gb/showGb? type = online&hcno = A495CA355BAF00D962AA8DD84C3B2C16.
[40] 屈宝强,王凯. 数据出版视角下的科学数据同行评议[J]. 图书馆杂志,2017,36(10):71 - 77.
[41] DATAVERSE. Data management plans[EB/OL]. [2018 - 11 - 15]. http://best-practices. dataverse. org/data-management/.

作者贡献说明:
秦顺:研究主题及论文框架确定,数据调研,撰写与修改论文;
汪全莉:修正研究思路,指导论文写作;
邢文明:提出论文修改建议,校对数据。

Research and Enlightenment on the Services of Open Access Publishing Platform for Scientific Data in Europe and America

Qin Shun Wang Quanli Xin Wenming
School of Public Administration, Xiangtan University, Xiangtan 411105

Abstract: [**Purpose/significance**] The value of data and the importance of open access publishing for scientific data has become a consensus. The experience on service construction of open access publishing platform for scientific data in Europe and America has a great significance for reference. [**Method/process**] This paper chose 14 typical open access publishing platforms for scientific data in Europe and America as samples, according to the data publishing life cycle of "collect-distribute-reuse", and discussed the service construction on five aspects: the policy or vision of scientific data publishing, scientific data consolidation, identification and interaction, scientific data publishing and distribution, scientific data reference, data life cycle management and publishing quality control, then summarized the characteristics and experience of its service construction. [**Result/conclusion**] The useful enlightenment on the service construction of open access publishing platform for scientific data in China was drawn up, including: formulating policies covering the whole life cycle of scientific data publishing, and attaching the importance of the key issues on scientific data publishing service which involves data consolidation, data identification, data reference and data review, etc.

Keywords: scientific data open access data publishing data identification data reference